Robust Linear Clustering †

L.A. García-Escudero
University of Valladolid, Valladolid, Spain
A. Gordaliza
University of Valladolid, Valladolid, Spain
R. San Martín
University of Valladolid, Valladolid, Spain
S. Van Aelst
Ghent University, Ghent, Belgium.
and R. Zamar
University of British Columbia, Vancouver, Canada.

Summary. Non-hierarchical clustering methods are frequently based on the idea of forming groups around "objects". The main exponent of this class of methods is the k-means method, where these objects are points. However, clusters in a data set may often be due to the existence of certain relationships among the measured variables. For instance, we can find linear structures such as straight lines, planes and so on, around which the observations are grouped in a natural way. These structures are not well represented by points. We present a method that searches for linear groups in the presence of outliers. The method is based on the idea of impartial trimming. We search for the "best" subsample containing a proportion $1-\alpha$ of the data and the best k affine subspaces fitting to those non-discarded observations by measuring discrepancies through orthogonal distances. The population version of the sample problem will also be considered. We prove the existence of solutions for the sample and population problems together with their consistency. A feasible algorithm for solving the sample problem is described as well. Finally, some examples showing how the proposed method works in practice are provided.

Keywords: Robustness, Trimming, Affine Subspaces, Principal Components, Orthogonal Regression, Trimmed k-means.

1. Introduction

Non-hierarchical methods in Cluster Analysis are usually based on the idea of forming groups around "centers", which represent the typical behavior of the points in each group. Clustering is an important tool for unsupervised learning that has received a lot of attention in the literature. Many clustering methods and algorithms have been proposed in various fields such as statistics (see e.g. Hartigan 1975, Kaufman and Rousseeuw 1990, Banfield and Raftery 1993, Scott 1992, Silverman 1986), data mining (see e.g. Ng and Han 1994, Zhang et al. 1997, Bradley et al. 1998, Murtagh 2002), machine learning (see e.g. Fisher 1987), and pattern recognition (see e.g. Duda et al. 2000, Fukunaga 1990). The main exponent of this class of methods is the k-means method (McQueen 1967 and Hartigan and Wong 1979), based on the Least Squares criterion from which it inherits a great drawback: its lack of robustness. In order to solve the lack of robustness of the k-means method, Cuesta-Albertos et al. (1997) introduced the trimmed k-means method which allows that a proportion α of (possible) outlying observations is left unassigned to the resulting groups.

The presence of clusters in a data set is sometimes due to the existence of certain relationships among the measured variables, which may adopt different patterns in each group. For instance, we can find in a data set several linear structures such as straight lines, planes and so on, around which the observations could be grouped in a natural way. Hosmer (1974), Lenstra et al (1982) and Späth (1982) made first attempts of clustering in this type of data sets by fitting a mixture of two simple linear regression models. Alternative

†*Address for correspondence:* Luis A. García-Escudero, Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, 47011 Valladolid, Spain.
E-mail: lagarcia@eio.uva.es

algorithms for the two dimensional problem have been introduced by Murtagh and Raftery (1984) and Phillips and Rosenfeld (1988). Desarbo and Cron (1988) stated the problem in general dimensions and for an arbitrary number of linear groups. They used Maximum Likelihood Estimation and the EM algorithm to solve that problem. Alternative solutions have been proposed by Desarbo et al. (1989), Kamgar-Parsi et al. (1990) and Peña et al. (2003). Hennig (2003) studied different models that yield linear clusters through linear regression.

Recently, Van Aelst et al. (2006) addressed the problem of linear grouping by using an orthogonal regression approach and obtained a very good performance in several problems where no outlying observations were present. However, this approach suffers from a serious lack of robustness problem. Note that it reduces to the classical non-robust Principal Components when we search for only one group. Some potential applications in fields like computer vision (see e.g. Stewart 1999), pattern recognition (see e.g. Murtagh and Raftery 1984 and Campbell et al. 1997) or tomography (Maitra 2001) suggest that more attention should be paid to robustness, because the presence of noise in the data sets may be frequent in all these fields of application. Clustering around lines in presence of noisy data has been previously treated in Banfield and Raftery (1993) and Dasgupta and Raftery (1998) (by considering mixture fitting where noise is modeled through a uniform component of the mixture) and in Chen et al. (2001) and Müller and Garlipp (2005) (by considering redescending M-estimators and following an approach closely related with nonparametric density estimation techniques). Agostinelli and Pellizzari (2006) proposes a hierarchical clustering approach based on iterated Least Quantile Squares regressions.

We present a new method that searches for linear groups in the presence of outliers by robustifying the orthogonal regression based linear grouping algorithm of Van Aelst et al. (2006). The method is based on the idea of "impartial trimming" (Gordaliza 1991 and Cuesta-Albertos et al. 1997). The key idea is that the data itself tell us which observations should be deleted. This approach, apart from allowing us to cluster around general d-dimensional subspaces (not only around straight lines), differs from the above mentioned methodologies in that it is based on a trimmed least squares criterium and, so, it incorporates the robustness in a very natural way.

Given a sample $\{x_1, ..., x_n\}$ of observations in $\mathbb{R}^p$, $0 \leq \alpha < 1$ (the expected proportion of outliers to be trimmed off), d (the dimension of the affine subspaces with $1 \leq d < p$) and $k \in \mathbb{N}$ (the number of groups that we are searching for), we look for the solution of the problem:

$$\min_{\mathbf{Y}\subset\{x_1,...,x_n\},\#\mathbf{Y}=[n(1-\alpha)]} \quad \min_{\{h_1,...,h_k\}\subset\mathcal{A}_d} \quad \frac{1}{[n(1-\alpha)]} \sum_{x_i\in\mathbf{Y}} \min_{j=1,...,k} \| x_i - \mathrm{Pr}_{h_j}(x_i) \|^2, \tag{1}$$

where $\mathcal{A}_d := \{h \subset \mathbb{R}^p, h \text{ is a } d\text{-dimensional affine subspace}\}$ and $\mathrm{Pr}_h(\cdot)$ denotes the orthogonal projection onto h.

Any solution $H^0 = \{h_1^0, ..., h_k^0\}$ of (1) induces a partition of the non-trimmed observations into k linear clusters in the following way: the cluster C_j consists of all observations in the sample which are closer to h_j^0 than to the remaining $k-1$ optimal subspaces in H^0.

If we assume $\{x_1, ..., x_n\}$ to be the realization of a random sample from a theoretical distribution P, the sample or empirical problem in expression (1) admits a theoretical or population counterpart that will be described in Section 2. Because of the existence of a population version of the original problem, the proposed method provides not only a tool for data analysis, but also estimates of some interesting population features. Existence of solutions for both, the sample and the population problems, will be shown. Moreover, the consistency of the solutions of the sample problem toward the population solution will be derived.

A feasible algorithm for solving the sample problem is presented as well. This algorithm combines ideas of the nonrobust linear grouping algorithm of Van Aelst et al. (2006) with techniques of the trimmed k-means algorithm in García-Escudero et al. (2003).

The case $k = 1$ provides a trimming-based robustification of Principal Components Analysis which is discussed in detail in Croux et al. (2007).

2. Population problem

Let P be an absolutely continuous probability distribution on $\mathbb{R}^p$, the population α-trimmed k affine subspaces problem for P is stated as follows:

Let $\alpha \in (0,1)$ and $k \in \mathbb{N}$, and, for every $H = \{h_1, ..., h_k\} \subset \mathcal{A}_d$ and every Borel set A such that $P(A) = 1 - \alpha$, we measure the k-variation around H given A by

$$V_A(H) := \frac{1}{1-\alpha} \int_A d(x, H)^2 dP(x).$$

with $d(x, H) = \min_{j=1,...,k} \|x - \Pr_{h_j}(x)\|$. Then:

(a) we obtain the k-variation given A, by minimizing over H:

$$V_A := \inf_{H \subset \mathcal{A}_d, \#H=k} V_A(H),$$

(b) and, finally, we obtain the α-trimmed k-variation by minimizing over A:

$$V_{k,\alpha} := \inf_{A: P(A)=1-\alpha} V_A.$$

By solving this double minimization problem, we achieve an optimal set A_0 and k optimal affine subspaces $H^0 = \{h_1^0, ..., h_k^0\}$ such that $V_{A_0}(H_0) = V_{k,\alpha}$.

As in Van Aelst et al. (2006), we use orthogonal distances to measure the discrepancies because we do not assume the existence of any privileged variable (that we want to explain in terms of the others). Note that the orthogonal distances between the observations and the hyperplanes are not scaled, so this objective function implicitly assumes equal variances along the linear subspaces. For sake of simplicity, we have stated the problem for absolutely continuous distributions assuming that Borel sets with probability exactly equal to $1 - \alpha$ do always exist. However, the problem can be stated more generally by introducing trimming functions (see Gordaliza 1991) that allow the partial participation of the points in the optimal set.

The next result provides a characterization of the optimal sets. Some notation will be needed: Given $h \in \mathcal{A}_d$ and a radius r, we define the "strip" $S(h, r)$ as $S(h, r) := \{x \in \mathbb{R}^p : \| x - \Pr_h(x) \| \leq r\}$. Analogously, for a set $H = \{h_1, ..., h_k\} \subset \mathcal{A}_d$, the "generalized strip" can be defined as $S(H, r) := \cup_{j=1}^k S(h_j, r) \equiv \{x \in \mathbb{R}^p : d(x, H) \leq r\}$. The following result tells us that the optimal sets are essentially generalized strips centered at some H and with radius

$$r_\alpha(H) := \inf\{r \geq 0 : P(S(H, r)) = 1 - \alpha\}.$$

PROPOSITION 1. *For every $H = \{h_1, ..., h_k\} \subset \mathcal{A}_d$, we have:*

(a) *$V_{S(H, r_\alpha(H))}(H) \leq V_A(H)$ for every Borel set A such that $P(A) = 1 - \alpha$.*

(b) *The inequality in* (a) *is strict if and only if $P(A \triangle S(H, r_\alpha(H))) > 0$ ($\triangle$ denotes the symmetric difference between sets).*

Proposition 1 allows us to simplify the original double minimization problem to (only) the optimal determination of a set H_0 of k optimal affine subspaces. Note that once H_0 has been determined, the optimal set A_0 is given by the generalized strip $A_0 = S(H_0, r_\alpha(H_0))$.

Moreover, in order to better characterize the optimal sets, we provide the following "self-consistency" result (see, e.g., Tarpey and Flury 1996). Let us consider the partition of the optimal set A_0 onto the k subsets:

$$C_j^0 = \{x \in \mathbb{R}^p : x \in S(H_0, r_\alpha(H_0)) \text{ and } \|x - \Pr_{h_j^0}(x)\| \leq \|x - \Pr_{h_l^0}(x)\| \text{ for } l \neq j\}, j = 1, ..., k,$$

and, $P_{C_j^0}$ denotes the conditional probability of P conditioned to be in C_j^0. Then:

PROPOSITION 2. *For P an absolutely continuous distribution with finite second order moments, if $H_0 = \{h_1^0, ..., h_k^0\}$ are the k optimal affine subspaces, then each h_j^0 must be an affine subspace spanned by the ordinary population principal components of the distribution $P_{C_j^0}$.*

Although a finite second order moment condition is assumed in Proposition 2, we will see that no moment conditions are needed to proof the existence of solutions or the consistency result. This lack of moment conditions is important because outliers are frequently modeled by heavy-tailed distributions. Note that the optimal trimmed k affine subspaces are defined as subspaces that minimize trimmed squared orthogonal loss functions instead of "principal components" based on covariance matrices. Proposition 2 says that the two views coincide when the covariance matrices do exist. This proposition also suggests us the application of an algorithm as described in Section 4 which can be seen as a kind of "(trimmed) self-consistency" algorithm (Tarpey 1999).

The next result establishes the existence of a solution for the previously stated problem, without assuming the existence of moments:

THEOREM 1. *Let P be an absolutely continuous probability distribution on $\mathbb{R}^p$ and $\alpha \in (0,1)$, then there always exist a set A_0 and a set of k affine subspaces H^0 such that $V_{A_0}(H_0) = V_{k,\alpha}$.*

The proof of this result is deferred to the appendix. This proof requires some technical lemmas including an interesting "continuity" result (Lemma 2) and a result telling us that the objective function $V_{k,\alpha}$ decreases when the number of groups k is increased (Lemma 3).

3. Sample problem and consistency

If $\{X_n\}_n$ is a sequence of independent, identically distributed (i.i.d.) random vectors sampled from the distribution P, the empirical distribution is defined as $P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i)$. The original problem stated in (1) follows by considering the same problem as in Section 2 but replacing the (unknown) underlying distribution P by the empirical distribution P_n.

Although the existence result was stated for absolutely continuous distributions, the existence of solutions in the empirical case can be easily derived. Note that there exists a finite number of ways to split $\{x_1, ..., x_n\}$ into k groups such that its total number of elements is $[n(1-\alpha)]$. Then, for each partition, the optimal k affine subspaces are obtained by resorting to orthogonal regression of the observations in the groups. This yields a finite number of candidate k affine subspaces from which the optimal solution needs to be selected.

In this section, we provide a consistency result stating the convergence of the sample solutions to the population solution. The convergence between affine subspaces here must be seen as the convergence of the distances to the origin and the possible choice of converging sequences of unitary spanning vectors.

THEOREM 2. *Let $\{X_n\}_n$ be a sequence of i.i.d. random vectors with common absolutely continuous distribution P such that its associated (population) problem admits a unique solution H_0. If $\{H_n\}_n$ is a sequence of sample k optimal affine subspaces and $\{V^n_{k,\alpha}\}_n$ is the associated sequence of sample empirical α-trimmed k-variations, then the convergences in probability $H_n \to H_0$ and $V^n_{k,\alpha} \to V_{k,\alpha}$ hold.*

The uniqueness of the solution can not be guaranteed for general probability distributions P. There exists a uniqueness result in the $k = 1$ case for unimodal elliptical distributions when their d largest eigenvalues are bigger than the $p-d$ smallest ones (Croux et al. 2007). Unfortunately, it is difficult to extend this result to the general $k > 1$ case.

4. Algorithm

The computation of the optimal empirical α-trimmed k affine subspaces has obviously a high computational complexity, because a search in the combinatorial space of subsets of a given data set is needed. Hence, exact algorithms are, in general, not feasible and the development of an adequate approximate algorithm is as important as the procedure itself.

The algorithm introduced here is an adaptation of the one proposed for computing the empirical trimmed k-means (García-Escudero et al. 2003). This last algorithm may be seen as a combination of the classical k-means algorithm and the rationale behind the FAST-MCD algorithm in Rousseeuw and van Driessen (1999) for computing the Minimum Covariance Determinant (MCD) estimator. In trimmed k-means a "concentration" step (or C-step) as in the fast-MCD algorithm was applied by keeping the $[n(1-\alpha)]$ observations with lowest Euclidean distances from their respective centers. Now, in this new set-up, we keep the $[n(1-\alpha)]$ observations with smallest orthogonal distances from the closest subspace among the k affine subspaces from

the previous iteration. Then, k new affine subspaces are obtained through orthogonal regression (i.e., solving k Principal Components problems) as in the linear grouping algorithm of Van Aelst et al. (2006). Thus, for a given data set $\{x_1, ..., x_n\}$, a fixed number of groups k, and a fixed trimming fraction α, the algorithm can be described as follows:

Step 1: We first scale the variables to avoid numerical accuracy problems. Each variable is scaled robustly by dividing through its median absolute deviation.
Step 2: Randomly select k starting affine subspaces in $\mathcal{A}_d$ (for instance, draw at random $(d+1) \times k$ observations in general position from the whole data set and use them to obtain k affine subspaces where each one is determined by $d+1$ points). Note that each hyperplane is determined by the mean x_0^j of the $d+1$ points and a matrix U_0^j whose columns are the d unitary eigenvectors corresponding to the nonzero eigenvalues of the sample covariance matrix of the $d+1$ observations.
Step 3: The "Concentration" step:
Assume that $H = \{h_1, ..., h_k\} \subset \mathcal{A}_d$ are the k affine subspaces obtained in the previous iteration:

Step 3.1: Compute the distances $d_i = d(x_i, H)$, $i = 1, ..., n$, between each observation and its closest subspace among the k affine subspaces from the previous iteration. Determine the set C that consists of the $[n(1-\alpha)]$ observations with lowest d_i's where

$$d_i^2 = \inf_{j=1,...,k} \left\| (I - U_0^j (U_0^j)') (x_i - x_0^j) \right\|^2.$$

Step 3.2: Partition C into $C = \{C_1, ..., C_k\}$ where the points in C_j are those observations closer to h_j than to any of the other affine subspaces h_l with $l \neq j$. That is, $C_j := \{x_i \in C : d(x_i, h_j)^2 = d_i\}$.
Step 3.3: Let x_1^j be the sample mean of the observations in C_j and U_1^j be a matrix containing the d largest orthogonal unitary eigenvectors of the sample covariance matrix of the observations in C_j. The k affine subspaces $H = \{h_1, ..., h_k\}$ for the next iteration will be the k affine subspaces passing through x_1^j and spanned by the vectors given by the columns of U_1^j, for $j = 1, ..., k$.

Step 4: Repeat the "concentration"-step a few (e.g. 10) times. After these iterations, compute the final evaluation function

$$\frac{1}{[n(1-\alpha)]} \sum_{j=1}^{k} \sum_{x_i \in C_j} d_i^2. \tag{2}$$

Step 5: Draw random starting subspaces (i.e., start from step 1) several times (e.g. 500 times), keep the solutions (e.g. 10) leading to minimal values of the evaluation function (2) and fully iterate those to choose the optimal solution.

Note that the algorithm reduces to the Linear Grouping Algorithm in Van Aelst et al. (2006) when $\alpha = 0$. For each random start, the iterative procedure in Step 3 converges to a locally optimal solution. As argued in Rousseeuw and Van Driessen (1999), a few concentration steps usually suffice to decide which random starts converge to a good global solution. On the other hand, a sufficient number of random starts is needed to have high enough probability that at least one random start converges to the global solution. Similarly as in Van Aelst et al. (2006), one could calculate the minimal number of starting values that is needed to have 95% probability of obtaining at least one starting solution that is optimal in the sense that it is outlier free and contains exactly $d+1$ observations of each of the k groups. However, this number depends on k, the relative sizes of the k groups, α, d and p. In practice, not all this information will be available beforehand. Moreover, the resulting number of random starts is much higher than necessary, because the concentration steps in Step 3 allow the algorithm to converge to a good solution from any reasonable initial random start. In our experience, taking 500 or 1000 random starts is sufficient to find a good solution.

REMARK 1. *Although our algorithm is consistent for the population partition induced by the trimmed mean squared criterium, this partition is not necessarily the most interesting partition in all applications. Note that, without trimming, the proposed algorithm can be viewed as a classification-likelihood EM-algorithm (see, e.g., Celeux and Govaert 1992) which is known to be inconsistent for estimating the underlying mixture model parameters (see Bryant and Williamson 1978). Therefore, it is possible that clusters generated by a mixture distribution may not be uncovered by our algorithm. An ideal situation for our algorithm would be,*

for instance, a population consisting of linear d-dimensional subspaces plus $(p-d)$-dimensional spherical and equally scattered Gaussian error terms lying in the orthogonal complement of these linear subspaces (see also García-Escudero et al. 2008).

5. Examples

In this section we illustrate the performance of the proposed approach on simulated and real data.

5.1. Simulated examples

We consider synthetic datasets generated according to the slanted π configuration (random points from three linear models in two dimensions) as already used in Van Aelst et al. (2006) but we add different types of outliers to illustrate the robustness of the trimmed affine subspaces. In Figure 1(a) we generated $n = 300$ points according to the slanted π configuration, but we replaced 50 points by scattered outliers (marked $\star$).

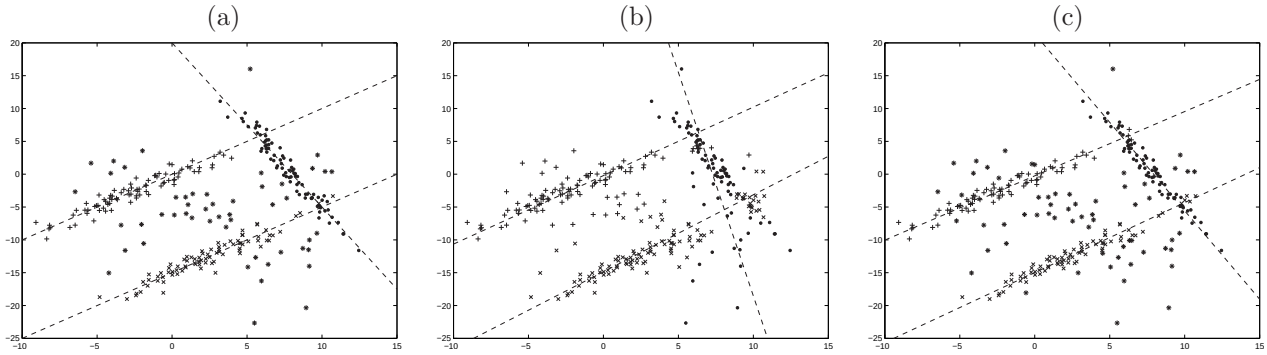

Fig. 1. (a) Slanted π data set of size $n = 300$ with 50 outliers; (b) LGA solution ($\alpha = 0\%$) for $k = 3$ groups; (c) Robust solution ($\alpha = 20\%$) for $k = 3$ groups

Note that we have both outliers that are far from the bulk of the data, as well as *inliers* that are not close to any of the linear patterns but do belong to the bulk of the data because they lie between the linear patterns. Figure 1(b) illustrates the nonrobustness of the linear grouping algorithm (LGA) as proposed in Van Aelst et al. (2006). In this example, the outliers mainly affected the line at the top of the π. Moreover, the residual variability has become high because all outliers have been assigned to their closest line. On the other hand, if we apply the robust linear grouping algorithm with 20% trimming then we obtain the result in Figure 1(c) where the trimmed points are now marked with $\star$. Comparing this result with Figure 1(a) reveals that we now have successfully retrieved the linear patterns and that the method trimmed all the outliers. Note that also some points that actually lie close to a hyperplane have been trimmed as a consequence of the choice of a large trimming fraction α. The trimming fraction has been taken larger than necessary to mimic the use of the procedure in practice where the fraction of outliers is unknown. However, by comparing the distance d_i between a trimmed point and its closest hyperplane to the distances of the points assigned to that hyperplane, we can easily decide which points should be really trimmed and which can actually be assigned to a group. In this way the clustering can be further improved.

It is obvious that assignment of points is difficult in the intersection regions between two (or more) hyperplanes and errors will be inevitable. Note that in practice the true group membership is unknown. Points in these 'intermediate' regions will be close to more than one hyperplane and could be given double (or multiple) membership. To measure how strongly each object belongs to its assigned group, Van Aelst et al. (2006) extended the silhouette width (Rousseeuw 1987) to the linear grouping setting. The silhouette width compares the distance of an object to its assigned group with the distance to its neighbor (the second closest hyperplane). The larger the silhouette width of an object, the more confident one can be about the correctness of its assignment. On the other hand, objects with smaller silhouette widths are more likely to be assigned incorrectly. Alternatively, posterior probabilities and Bayesian factors can be used to measure strength of group membership if a model is used for each of the linear groups (see Van Aelst et al., 2006).

The next two examples consider extreme situations. In Figure 2(a) we have 100 (33%) points that are scattered outliers, which makes it hard to even detect the linear patterns by eye if the symbol coding would

be removed. Figure 2(b) contains a tight cluster of inliers (50 points), which can be identified easily by eye, but because it is so tight, it causes many problems for the nonrobust LGA. In both cases the LGA solution becomes unstable and completely misses at least one of the three linear patterns as shown in Figures 2(c) and (d). On the other hand, even in such extreme cases, the robust linear grouping algorithm can still identify the linear patterns as can be seen from Figures 2(e) (40% trimming) and (f) (25% trimming). Note that we verified that the three linear models shown in Figures 2(a) and (b) correspond to the population solutions of the α-trimmed linear grouping problem. Therefore, it is very likely that the robust linear grouping solutions shown in Figures 2(e) and (f) are in fact global solutions. These extreme examples show the powerful performance of the robust linear grouping algorithm to detect linear patterns in the presence of contamination.

5.2. Corridor-walls recognition

Computer Vision is an interesting field where linear grouping methods can be applied. Moreover, the approach developed here is especially appealing in computer vision due to the different sources of noise often present in this context. Note that robust estimation methods related to trimming have been adapted before for computer vision applications (see e.g. Meer et al. 1991, Jolion et al. 1991, and Stewart 1995).

In some applications of Computer Vision, a set of two or three dimensional measurements is taken from a place and we must try to recognize different structures in the data to help us identify the objects present in the room. In our example, a laser device is introduced in a corridor of an office building and we want to recognize the main elements constituting the corridor. The device throws a laser ray which touches a point from the object found at the end of its trajectory and takes a three dimensional measurement of the placement of that point with respect to a fixed reference system. The laser device sweeps all possible directions following a dense grid of solid angles and it generates a large three dimensional data set. The goal is to recognize the exact position of the walls and the ceiling of the corridor, but we could have some noise due to objects that were placed on the floor or attached to the walls or the ceiling. To make the figures interpretable, we have selected only a small part of the whole data set, but the performance of our method is good even in more complex situations than the one presented here.

Figure 3(a) shows the data set we want to analyze. In this figure we can easily guess the linear structures (planes) corresponding to the two walls and the ceiling. We can also see some points corresponding to an object lying on the floor. When we apply the robust linear grouping algorithm with $k = 3$ and $\alpha = .15$, we obtain the three clusters shown in Figure 3.

Figure 4(a) shows the trimmed points, which are the points placed farthest away from the linear structures we have found. Figure 4(b) shows the distances of the trimmed points to the corresponding planes. We can see that the trimmed points come from three different sources. The group with the largest distances corresponds to the object placed on the floor, which is far away from all the planes. The group of trimmed points with "log(distances+1)" around 0.07 corresponds to a heating radiator hanging on the left wall. Finally, there exists a third group of trimmed points whose distances are quite close to the optimal radius which served as the cutoff point to decide whether an observation should be trimmed off (if its distance exceeds that radius) or not. Note that the choice $\alpha = .15$ was rather subjective but the relative "proximity" of the distances to the optimal radius could be used to further decide whether a trimmed data point can be recovered as a "regular" constituent of the walls or it merely corresponds to some irregularities (caused perhaps by inexperienced or unprofessional building workers). For instance, a more detailed analysis of the original data has shown a not very high finish in the ceiling and in some corners of this corridor together with some small damaged zones in the walls that this method was able to detect.

Note that the consideration of a high trimming level α followed by a careful examination of a plot like that appearing in Figure 4(b) could in general be a good strategy with this kind of data. We would also like to stress that, due to its linear shape, we could even recover the radiator by setting $k = 4$ in the procedure. To make our method a useful tool in Computer Vision data driven procedures for automatically doing the examination of plots like that in Figure 4 will be needed.

6. Discussion

We introduced a robust method to detect linear structures in a data set. Our method robustifies the linear grouping technique of Van Aelst et al. (2006) by using impartial trimming. We have shown that solutions of our method exist both at the sample and population level and moreover, the solution is consistent.

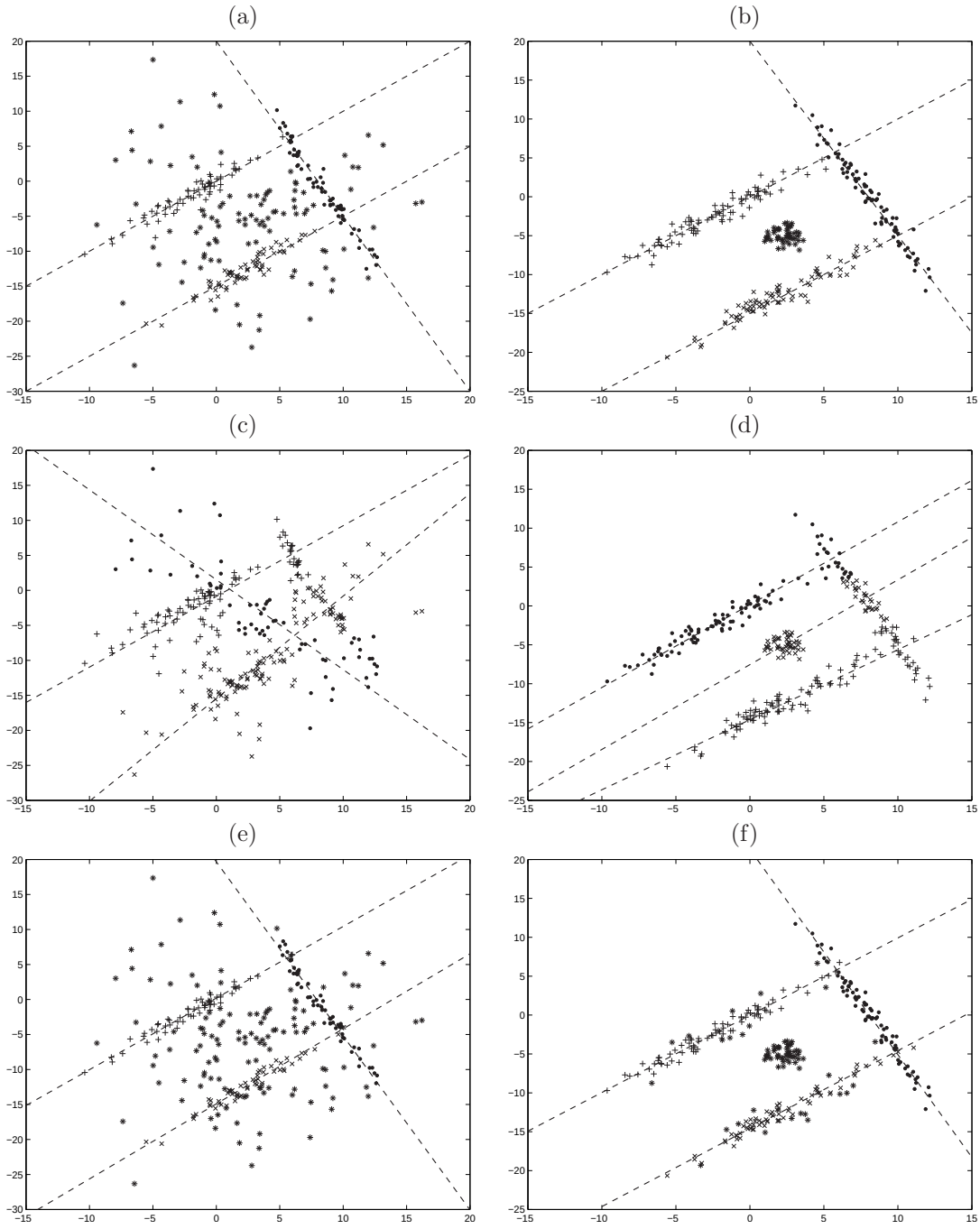

Fig. 2. (a) Slanted π data set of size $n = 300$ with 100 scattered outliers; (b) Slanted π data set of size $n = 300$ with a cluster of 50 inliers; (c) LGA solution ($\alpha = 0\%$) corresponding to (a) for $k = 3$ groups; (d) LGA solution ($\alpha = 0\%$) corresponding to (b) for $k = 3$ groups; (e) Robust solution ($\alpha = 40\%$) corresponding to (a) for $k = 3$ groups; (f) Robust solution ($\alpha = 25\%$) corresponding to (b) for $k = 3$ groups

We presented a computationally feasible algorithm based on concentration steps, that provides an adequate approximate solution to the problem. The examples have illustrated the usefulness of our method in practice. Note that this procedure based on orthogonal regression does not require the specification of a response variable.

Banfield and Raftery (1993) and Dasgupta and Raftery (1998) also considered the problem of detecting linear clusters in noisy data. These procedures use a model consisting of a mixture distribution where the noise is assumed to arise from a Poisson process with constant intensity. Clearly, the performance of the

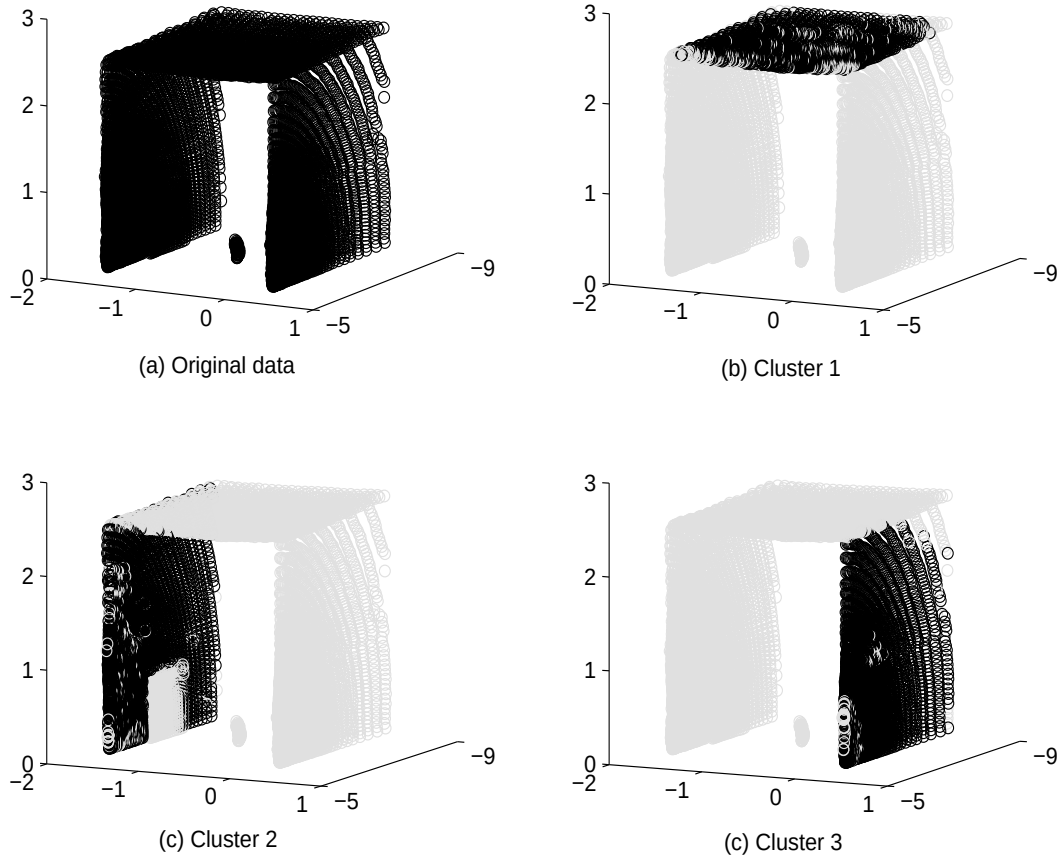

Fig. 3. Results of the robust linear clustering algorithm for the "corridor-walls" data set when $k = 3$ and $\alpha = .15$.

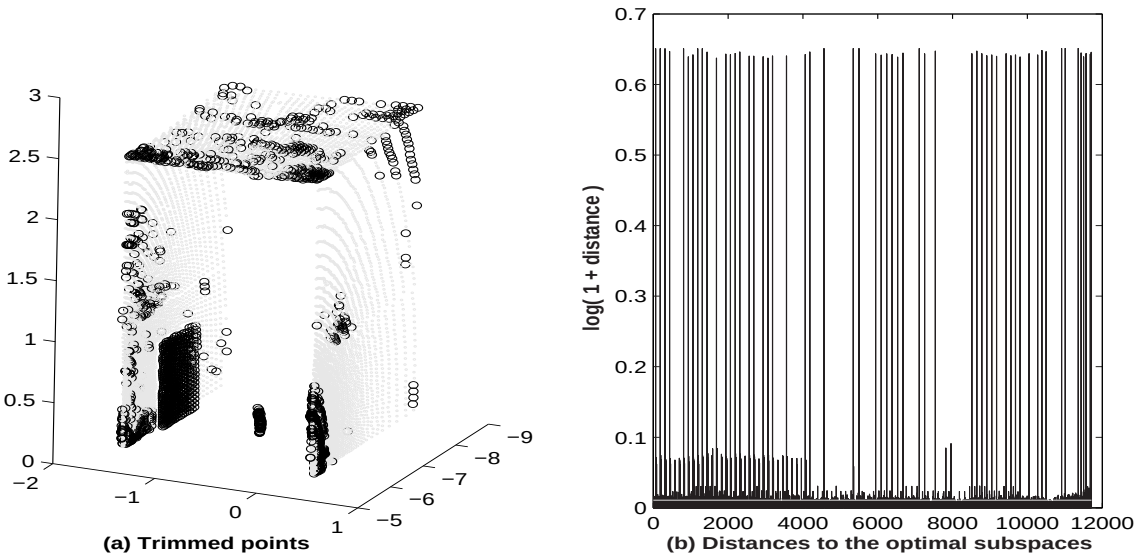

Fig. 4. (a) Trimmed observations for the "corridor-walls" data set ($k = 3$ and $\alpha = .15$). (b) Distances of the observations to their closest optimal subspaces. Observations with distances greater than the optimal radius (horizontal white line) were the trimmed ones.

methods strongly dependents on the validity of the model assumptions. In the context of nonrobust linear clustering, Van Aelst et al. (2006) showed that MCLUST (Banfield and Raftery 1993) cannot always detect linear patterns that can be found by the linear grouping algorithm. The procedure of Chen et al. (2001) is limited to detecting lines in two-dimensional data. Müller and Garlipp (2005) considered a procedure based on local minima of orthogonal regression redescending M-estimators. Note that the algorithm requires good starting values. Our method can be used to find such starting values in the presence of outliers. Agostinelli and Pellizzari (2006) proposed a hierarchical clustering approach based on robust orthogonal regression through iterated Least Quantile Squares regressions. However, properties of this method such as consistency and behavior in the presence of outliers have not been investigated yet.

Our technique requires the input of two tuning parameters, the trimming fraction, α and the number of linear structures, k. These two parameters are often related. If the dataset is not suspected to contain a large fraction of outliers, a trimming fraction between 0.05 and 0.15 would be recommended. On the other hand, if the trimming fraction is too small, then the linear grouping may be distorted by the outliers, leading to an incorrect grouping. In our experience, the linear grouping procedure can often detect the 'core' of the linear groups even when a trimming fraction larger than necessary is being used. Therefore, for noisy data, a large enough trimming fraction in the beginning, e.g. α in the range 0.25-0.35 is recommended to reliably detect the linear structures in the data without adverse effects of the outliers. However, an entire cluster may be trimmed and therefore careful examination of the clustering results and trimmed points is necessary when using a large trimming fraction. By examination of the distances d_i of the trimmed points (e.g. using a graphical representation as in Figure 4) it can then be checked whether points that are close to a hyperplane have been trimmed. In such case these points can be assigned to their closest hyperplanes. The remaining trimmed points can be conveniently color tagged and graphically examined to determine whether they are isolated outliers or a cluster. Examination of high dimensional data can be done with the help of high level graphical tools such as dynamic projections. Dynamic projections have been successfully implemented in recent years by software such as XGobi (Swayne et al. 1998) and its successor GGobi (Swayne et al. 2003). They can be powerful in showing high-dimensional data structure, including the structure of outliers. The so-called "grand tour" provides an overview of the data through a random continuous sequence of 2D projections (1D or 3D projections have also been proposed). Alternatively, other graphical techniques specially aimed to clustering problems may be used (see Hennig and Christlieb 2002). The graphical procedures developed in Garcia-Escudero et al. (2003) can be useful to select the number of groups. Moreover, to check whether a group has been completely trimmed, it can be instructive to compare the current solution to the solution that is obtained when the number of groups, k, is increased by one and the trimming fraction is taken lower. If the problem at hand does not suggest any reasonable values for k, then graphical procedures as developed in García-Escudero et al. (2003) can be very useful as well to select the number of groups. If further linear structures exist among the trimmed points, or a substantial number of trimmed observations can be assigned to existing linear structures, then the analysis can be re-run with adjusted values of the trimming fraction α and/or the number of linear groups, k.

Our procedure detects subspaces with the same dimension d in a p-dimensional data set. In practice, subgroups of different dimensions can exist in a p-dimensional data set. For example, a two-dimensional data set can contain linear structures (dimension 1) as well as point clusters (dimension 0). However, our technique can be used as the basis of a multistage procedure as outlined in Van Aelst et al. (2006) where each of the $p-1$ dimensional subspaces that has been detected is investigated further to determine whether it is a genuine homogeneous $p-1$ dimensional subgroup or whether it is a mixture of one or more lower dimensional subgroups.

Since the "self-consistency" property plays a key role in our approach, it seems natural to try extending it to the case of robust clustering around principal curves (Hastie and Stuetzle 1989). Clustering around principal curves has already been proposed (see, e.g., Banfield and Raftery 1992) and providing some robustness to these procedures is appealing. Stanford and Raftery (2000) handled the presence of background noise by modeling it through a uniform noise mixture component. However, one could also consider a trimming approach by allowing a proportion α of observations to be discarded. This is ongoing work.

7. Appendix: Proofs

7.1. Proof of Proposition 1

Let $S = S(H, r_\alpha(H))$ and a Borel set A such that $P(A) = 1-\alpha$. Note that $P(A \cap S^c) = P(A^c \cap S)$, since $1-\alpha = P(A \cap S) + P(A \cap S^c) = P(A \cap S) + P(A^c \cap S)$. Thus,

$$\begin{aligned}
\int_A d(x,H)^2 dP(x) &= \int_{A\cap S} d(x,H)^2 dP(x) + \int_{A\cap S^c} d(x,H)^2 dP(x) \\
&\geq \int_{A\cap S} d(x,H)^2 dP(x) + r_\alpha(H)^2 P(A\cap S^c) \\
&= \int_{A\cap S} d(x,H)^2 dP(x) + r_\alpha(H)^2 P(A^c\cap S) \\
&\geq \int_{A\cap S} d(x,H)^2 dP(x) + \int_{A\cap S^c} d(x,H)^2 dP(x) = \int_S d(x,H)^2 dP(x)
\end{aligned}$$

Note that previous inequalities are strict ones whenever $P(A \triangle S) > 0$ (and, consequently, $P(A \cap S^c)$ and $P(A^c \cap S)$ are so strictly positive). □

7.2. Proof of Proposition 2

If that result did not hold, we could strictly diminish the variation by replacing h_j^0 by the affine subspace spanned by the ordinary principal components of the probability distribution $P_{C_j^0}$ and, thus, H_0 would not be the optimal affine subspaces. □

In order to prove the existence result, three previous technical lemmas will be needed:

LEMMA 1. *For any $0 < \alpha < 1$, we have $V_{k,\alpha} < \infty$.*

Proof: Let us consider $\tilde{h} \in \mathcal{A}_d$, the affine subspace spanned by the origin and the first d vectors of the canonical basis in $\mathbb{R}^p$ and H equal to $\tilde{h}$ plus other $k-1$ different affine subspaces. Take $r > 0$ such that $P(S(H,r)) = 1-\alpha$, we easily see that $V_{k,\alpha} \leq r^2 < \infty$. □

For the following results, we introduce the α-trimmed variation around H defined as:

$$V_\alpha(H) := \frac{1}{1-\alpha} \int_{S(H, r_\alpha(H))} d(x,H)^2 dP(x).$$

LEMMA 2. *Let $H_n = \{h_1^n, ..., h_k^n\} \subset \mathcal{A}_d$ be a sequence of sets of affine subspaces, $n = 0, 1, 2, ...$, such that $H_n \to H_0$ then $V_\alpha(H_n) \to V_\alpha(H_0)$ as $n \to \infty$.*

Proof: Let $r_n = r_\alpha(H_n)$ and $S_n = S(H_n, r_n)$, $n = 0, 1, 2, ...$, and, $D_n(\cdot) = I_{S_n}(\cdot) d(\cdot, H_n)^2 - I_{S_0}(\cdot) d(\cdot, H_0)^2$. We have

$$\begin{aligned}
(1-\alpha)(V_\alpha(H_n) - V_\alpha(H_0)) &= \int_{E_n} D_n(x) dP(x) + \int_{F_n} D_n(x) dP(x) + \int_{G_n} D_n(x) dP(x) \\
&:= A_n^{(1)} + A_n^{(2)} + A_n^{(3)},
\end{aligned}$$

with $E_n = S_0^c \cap S_n$, $F_n = S_0 \cap S_n^c$ and $G_n = S_0 \cap S_n$ (notice that $D_n(x) = 0$ for every $x \in S_0^c \cap S_n^c$).

The sequence $\{r_n\}_n$ is clearly bounded (because $H_n \to H_0$) and $E_n \downarrow \emptyset$. Hence,

$$\begin{aligned}
\left|A_n^{(1)}\right| &\leq \left| \int_{E_n} I_{S_n}(x) d(x,H_n)^2 dP(x) \right| + \left| \int_{E_n} I_{S_0}(x) d(x,H_0)^2 dP(x) \right| \\
&\leq (r_n^2 + r_0^2) P(E_n) \to 0 \text{ as } n \to \infty.
\end{aligned}$$

In a similar way we can prove that $\left|A_n^{(2)}\right|$ converges to 0.

To study the convergence of $\left|A_n^{(3)}\right|$, let us consider:

$$\begin{aligned}
|A_n^{(3)}| &\leq \left| \int_{G_n} I_{S_n}(x) (d(x,H_n)^2 - d(x,H_0)^2) dP(x) \right| + \left| \int_{G_n} (I_{S_n}(x) - I_{S_0}(x)) d(x,H_0)^2 dP(x) \right| \\
&:= B_n^{(1)} + B_n^{(2)}.
\end{aligned}$$

But $B_n^{(2)} = 0$ because $I_{S_n}(x) = I_{S_0}(x) = 1$ for $x \in G_n$. For $B_n^{(1)}$, we have

$$B_n^{(1)} \leq \int_{G_n} \left| d(x,H_n)^2 - d(x,H_0)^2 \right| dP(x) \leq \sup_{x \in G_n} \left| d(x,H_n)^2 - d(x,H_0)^2 \right| P(G_n).$$

This last term converges to 0, because $P(G_n) \leq 1-\alpha$ together with the fact that $d(x,H_n)^2 - d(x,H_0)^2 \to 0$ pointwise and taking into account the uniform continuity of the real valued quadratic function $g(x) = x^2$ on the compact set $[0,T]$ with $T = \sup\{r_n\}_{n=0}^{\infty} < \infty$. □

LEMMA 3. *Let $H = \{h_1, ..., h_k\} \subset \mathcal{A}_d$ and $\alpha \in (0,1)$. The following statements are equivalent:*
(a) $V_\alpha(H) > 0$
(b) *there exists $h_0 \in \mathcal{A}_d$ such that $V_\alpha(H \cup h_0) < V_\alpha(H)$.*

Proof: We only prove that (a) implies (b), because the other implication is obvious.

Suppose that $V_\alpha(H) > 0$. Then we have $r_\alpha(H) > 0$ and $P(H) < 1-\alpha$. Moreover, for every $r < r_\alpha(H)$, we have that $P(S(H,r)) < 1-\alpha$. For every set $H \subset \mathcal{A}_d$, let us consider $S_H = S(H, r_\alpha(H))$. We can easily see that there exists an $m_0 \in \mathbb{R}^p$ and $r_0 > 0$ such that $B_0 = B(m_0, r_0)$ satisfies: (i) $P(B_0 \cap S_H) > 0$, (ii) $d(m_0, \{\mathrm{Pr}_{h_1}(m_0), ..., \mathrm{Pr}_{h_k}(m_0)\}) > 2/3 \cdot r_\alpha(H)$, and (iii) $r_0 < 1/3 \cdot r_\alpha(H)$.

Let $h_0 \in \mathcal{A}_d$ and such that $m_0 \in h_0$. We have:

$$\begin{aligned}
(1-\alpha)V_\alpha(H) &= \int_{B_0 \cap S_H} d(x,H)^2 dP(x) + \int_{B_0^c \cap S_H} d(x,H)^2 dP(x) \\
&> \int_{B_0 \cap S_H} d(x,h_0)^2 dP(x) + \int_{B_0^c \cap S_H} d(x,H)^2 dP(x) \qquad (3) \\
&\geq \int_{S_H} \min\{d(x,h_0), d(x,H)\}^2 dP(x) \\
&= \int_{S_H} d(x, H \cup h_0)^2 dP(x) \geq \int_{S_{H \cup h_0}} d(x, H \cup h_0)^2 dP(x) \qquad (4) \\
&= (1-\alpha)V_\alpha(H \cup h_0).
\end{aligned}$$

We have applied (i), (ii) and (iii) in order to get the strict inequality in (3). To achieve (4), we take into account that

$$\int_{S_H \cap S^c_{H \cup h_0}} d(x, H \cup h_0) dP(x) \geq \int_{S^c_H \cap S_{H \cup h_0}} d(x, H \cup h_0) dP(x)$$

because $P(S_H \cap S^c_{H \cup h_0}) = P(S^c_H \cap S_{H \cup h_0})$ and $d(x, H \cup h_0) \geq d(y, H \cup h_0)$ for every $x \in S_H \cap S^c_{H \cup h_0}$ and $y \in S^c_H \cap S_{H \cup h_0}$. □

7.3. Proof of Theorem 1

Recall that Proposition 1 tell us that $V_{k,\alpha} = \inf_{H \in \mathcal{A}_d, \#H = k} V_\alpha(H)$. Now, from Lemma 1, we can take a sequence of sets $H_n = \{h_1^n, ..., h_k^n\} \subset \mathcal{A}_d$ such that $V_\alpha(H_n) \downarrow V_{k,\alpha}$ as $n \to \infty$. First, we will prove the existence of a convergent subsequence of $\{H_n\}_n$ and, second, we will show that the limit set is an optimal k affine subspace.

If $d_n = \min_{j=1,...,k} d(h_j^n, 0)$, $r_n = r_\alpha(H_n)$, and $S_n = S(H_n, r_n)$, we can show that $\{d_n\}_n$ and $\{r_n\}_n$ are bounded sequences. Take $R > 0$ such that $P(B(0,R)) > \max\{1-\alpha, \alpha\}$. As $P(S_n) = 1-\alpha$, we trivially have $d_n - R \leq r_n \leq d_n + R$ for every $n \in \mathbb{N}$. Therefore, $\{r_n\}_n$ will be bounded if and only if $\{d_n\}_n$ is bounded. With this in mind, take two sequences of positive numbers $\{\xi_n\}_n$ and $\{R_n\}_n$ such that $\xi_n \downarrow 0$, $R_n \uparrow \infty$ and $P(B(0,R_n)) \geq 1-\xi_n$. If $\{d_n\}_n$ were not bounded, we could find a subsequence (denoted as the original one) with $d_n > 2R_n$ for every $n \in \mathbb{N}$. Then, we would have

$$V_\alpha(H_n) \geq \frac{1}{1-\alpha} R_n^2 \cdot P(S_n \cap B(0,R_n)^c) \geq R_n^2 \frac{1-\alpha-\xi_n}{1-\alpha} \uparrow \infty \text{ as } n \to \infty,$$

contradicting Lemma 1.

Thus, $\{d_n\}_n$ and $\{r_n\}_n$ are bounded, and, there exists a nonempty set $J \subseteq \{1, ..., k\}$ with

$$\begin{array}{l} \text{if } j \in J, \text{ then there exists a } d_j^0 \text{ such that } d_j^n \to d_j^0, \\ \text{if } j \notin J, \text{ then } d_j^n \to \infty \text{ as } n \to \infty \end{array} \tag{5}$$

(a subsequence denoted as the original one may be needed in (5)).

We can assume, without loss of generality, that $J = \{1, ..., m\}$ for $m \leq k$. We can trivially find some affine subspaces $h_j^0 \in \mathcal{A}_d$, verifying that $h_j^n \to h_j^0$ as $n \to \infty$ for $j \in J$, (because the distances to the origin and their unitary spanning vectors are bounded). Take, now, the sets $H_0^m = \{h_1^0, ..., h_m^0\}, H_n^m = \{h_1^n, ..., h_m^n\}$, and , $H_n^{-m} = \{h_{m+1}^n, ..., h_k^n\}$ and let $r'_n = r_\alpha(H_n^m)$ and $S'_n = S(H_n^m, r'_n)$. We have trivially that $r'_n \geq r_n$ and that $\{r'_n\}_n$ must also be a bounded sequence by a similar argument as before.

We can assume from (5), without loss of generality, that $d_j^n > 2R_n$ for $j \notin J$, $S(H_n^m, r_n) \cap S(H_n^{-m}, r_n) \cap B(0, R_n) = \emptyset$, and, $P(S(H_n^{-m}, r_n)) \leq \xi_n$, for every n. We, thus, have

$$(1-\alpha)V_\alpha(H_n^m) = \int_{B(0,R_n)\cap S'_n} d(x, H_n^m)^2 dP(x) + \int_{B(0,R_n)^c \cap S'_n} d(x, H_n^m)^2 dP(x) := C_n^{(1)} + C_n^{(2)}.$$

Note that

$$\begin{aligned} C_n^{(1)} &= \int_{B(0,R_n)\cap S(H_n^m, r_n)} I_{S'_n}(x) d(x, H_n^m)^2 dP(x) + \int_{B(0,R_n)\cap [S'_n - S(H_n^m, r_n)]} I_{S'_n}(x) d(x, H_n^m)^2 dP(x) \\ &\leq \int_{B(0,R_n)\cap S(H_n^m, r_n)} I_{S_n}(x) d(x, H_n^m)^2 dP(x) + (r'_n)^2 \cdot \xi_n, \end{aligned}$$

because $I_{S'_n}(x) = I_{S_n}(x)$ for all $x \in S(H_n^m, r_n)$ as $r'_n \geq r_n$ and $P\big(B(0, R_n) \cap [S'_n - S(H_n^m, r_n)]\big) \leq P(S(H_n^{-m}, r_n)) \leq \xi_n$. Therefore,

$$(1-\alpha)V_\alpha(H_n^m) \leq \int_{B(0,R_n)\cap S(H_n^m, r_n)} I_{S_n}(x) d(x, H_n^m)^2 dP(x) + 2(r'_n)^2 \xi_n$$

as, clearly, $C_n^{(2)} \leq (r'_n)^2 \cdot \xi_n$.

On the other hand, we have

$$(1-\alpha)V_\alpha(H_n) \geq \int_{B(0,R_n)\cap S(H_n^m, r_n)} I_{S_n}(x) d(x, H_n^m)^2 dP(x).$$

Thus, $(1-\alpha)V_\alpha(H_n) \geq (1-\alpha)V_\alpha(H_n^m) - 2(r'_n)^2\xi_n \geq (1-\alpha)V_{\alpha,m} - 2(r'_n)^2\xi_n$. But, as $\{r'_n\}_n$ is bounded sequence and $\xi_n \downarrow 0$, so, we get $2(r'_n)^2\xi_n \to 0$. Hence $V_{k,\alpha} = \lim_{n\to\infty} V_\alpha(H_n) \geq \lim_{n\to\infty} V_\alpha(H_n^m) \geq V_{m,\alpha}$.

Then, necessarily, $V_{k,\alpha} = V_{m,\alpha}$. Moreover, from Lemma 2, we have $\lim_{n\to\infty} V_\alpha(H_n^m) = V_\alpha(H_0^m)$ and it follows that $V_\alpha(H_0^m) = V_{m,\alpha}$, and then H_0^m will be optimal k affine subspaces. Now, if $m = k$, the proof is finished. Otherwise, if $m < k$, Lemma 3 implies that $V_\alpha(H_0^m) = 0$ and the existence is obviously guaranteed for $k \geq m$. □

7.4. Proof of Theorem 2

For some combinations of n and α there may not exist a set A with $P_n(A) = 1 - \alpha$. Therefore, for practical purposes, we use sets A containing $[n(1-\alpha)]$ observations in (1). In the proof of the consistency result, this implies also the consideration of a (asymptotically not important) term $O(n^{-1})$ which will be omitted.

It suffices to prove that every subsequence of $\{H_n\}_n$ (resp. $\{V_{k,\alpha}^n\}_n$) admits a new subsequence which converges to H_0 (resp. $V_{k,\alpha}$). We denote these subsequences (w.l.o.g.) as the original ones.

First, we show that $V_{k,\alpha}^n$ is uniformly bounded. To see this, just follow the same argument as in the proof of Lemma 1, but now we need the tightness of the empirical distribution sequence $\{P_n\}_n$ in order to guarantee the existence of a common radius r such that $P_n(S(\tilde{h}, r)) \geq 1 - \alpha$.

If $H_n = \{h_1^n, ..., h_k^n\}$, let $d_n = \min_{j=1,...,k} d(h_i^n, 0)$, $r_n = r_\alpha(H_n)$ and $S_n = S(H_n, r_n)$, $n = 0, 1, 2, ...$. We can also show that the sequences $\{d_n\}_n$ and $\{r_n\}_n$ are bounded. For doing this, we need again to use the tightness of $\{P_n\}_n$ for obtaining two sequences of positive numbers $\{\xi_n\}_n$ and $\{R_n\}_n$ such that $\xi_n \downarrow 0$,

$R_n \uparrow \infty$ and $P_n(B(0,R_n)) \geq 1-\xi_n$. Later, as we did in the proof of Theorem 1, we would see that if these sequences were not bounded, this would contradict the uniformly boundedness of $V^n_{k,\alpha}$.

Let r'_n and $S'_n = S(H_0, r'_n)$ such that $P_n(S'_n) = 1-\alpha$, $n = 0, 1, 2, \ldots$. The sequence $\{r'_n\}_n$ is again a bounded sequence and, so, we can assume that $r'_n \to r'_0$ for some $r'_0 > 0$.

The class $\{I_{S(H_0,r)}(\cdot) : r > 0\}$ is trivially a Glivenko-Cantelli class. Therefore,

$$o_P(1) = P_n(S'_n) - P(S'_n) = P_n(S'_n) - P(S'_0) + P(S'_0) - P(S'_n).$$

But $P(S'_n) - P(S'_0) = o_P(1)$ because $r'_n \to r'_0$ and the fact that P is an absolute continuous distribution.

Hence, $P(S'_0) = 1-\alpha$ and $r'_0 = r_0$. Moreover, the fact that $\{I_{S(H_0,r)}(\cdot)d(\cdot,H_0)^2 : r > 0\}$ is also a Glivenko-Cantelli class, the absolute continuity of P and the convergence $r'_n \to r_0$ imply

$$V^n_{k,\alpha} \leq \frac{1}{1-\alpha}\int_{S'_n} d(x,H_0)^2 dP_n(x) \to \frac{1}{1-\alpha}\int_{S_0} d(x,H_0)^2 dP$$

and, consequently,

$$\limsup_n V^n_{k,\alpha} \leq V_{k,\alpha}. \tag{6}$$

As $\{d_n\}_n$ and $\{r_n\}_n$ are bounded, there exists a nonempty set $J \subseteq \{1,\ldots,k\}$ and a subsequence of $\{H_n\}_n$ (denoted as the original one) such that: if $j \in J$, then there exist a $h_j \in \mathcal{A}_d$ such that $h^n_j \to h^0_j$, and if $j \notin J$, then $d^n_j \to \infty$ as $n \to \infty$. We can assume, without loss of generality, that $J = \{1,\ldots,m\}$ for $m \leq k$, and we use the notation: $H^m_0 = \{h^0_1,\ldots,h^0_m\}$, $H^m_n = \{h^n_1,\ldots,h^n_m\}$, and , $H^{-m}_n = \{h^n_{m+1},\ldots,h^n_k\}$.

As $\{r_n\}_n$ is bounded, we can assume that it admits a convergent subsequence (denoted as the original one) with limit, say, r. Then, $P_n(S(H^{-m}_n, r_n)) \to 0$, and, $P_n(S(H^m_n, r_n)) \to 1-\alpha$. Now, as $P_n(S(H^m_n, r_n)) \to P(S(H^m_0, r))$, we conclude that $P(S(H^m_0, r)) = 1-\alpha$. Furthermore, $\{I_{S(H,r)}(\cdot)d(\cdot,H)^2 : r > 0, H \subset \mathcal{A}_d \text{ and } \#H = k\}$ is also a Glivenko-Cantelli class of functions (its subgraph may be constructed from subgraphs of a finite dimensional family of functions). Thus, an argument almost equal to that applied in the proof of a Theorem 1 lead us to

$$\liminf_n V^n_{k,\alpha} \geq \frac{1}{1-\alpha}\int_{S(H^m_0,r)} d(x,H^m_0)^2 dP(x) \geq V_{m,\alpha}.$$

Therefore, by also applying (6), we have $V_{m,\alpha} = V_{k,\alpha} = \lim_n V^n_{k,\alpha}$. Finally, the absolute continuity of the distribution P together with the uniqueness of H_0 shows that $m = k$ and $H_0 = H^m_0$. □

Acknowledgements

This research of Luis Angel García-Escudero and Alfonso Gordaliza was partially supported by Ministerio de Educación y Ciencia and FEDER grant MTM2005-08519-C02-01 and by Consejería de Educación y Cultura de la Junta de Castilla y León grant PAPIJCL VA074/03. The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant P6/03 of the Belgian government (Belgian Science Policy). The research of Ruben Zamar was funded by NSERC.

References

Agostinelli, C. and Pellizzari, P. (2006) Hierarchical clustering by means of model grouping. In *From Data and Information Analysis to Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization* (eds. M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, W. Gaul), 246-253.

Banfield, J.D. and Raftery, A.E. (1992) Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Amer. Statist. Assoc.*, **87**, 7-16.

Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Bradley, P.S., Fayyad, U.M. and Reina, C.A. (1998) Scaling clustering algorithms to large databases. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.* AAAI Press, 9–15.

Bryant, P. and Williamson, J.A. (1978). Asymptotic behaviour of Classification Maximum Likelihood Estimates. *Biometrika* **65**, 273-281.

Campbell, J.G., Fraley, C., Murtagh, F. and Raftery, A.E. (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, **18**, 1539-1548.

Celeux, G. and Govaert, A. (1992). Classification EM algorithm for clustering and two stochastic versions. *Comput. Statit. Data Anal.,* **13**, 315-332.

Chen, H., Meer, P. and Tyler, D.E. (2001) Robust regression for data with multiple structures. In *2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001),* Vol. I, IEEE Computer Society, 1069-1075.

Croux, C., García-Escudero, L. A., Gordaliza, A. and San Marín, R. (2007) Robust Principal Components analysis based on trimming around affine subspaces. *Preprint.*

Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997) Trimmed k-means: An attempt to robustify quantizers. *Ann. Statist.,* **25**, 553-576.

Dasgupta, A. and Raftery, A.E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *J. Amer. Statist. Assoc.*, **93**, 294-302.

DeSarbo,W. and Cron, W. (1988) A maximum likelihood methodology for clusterwise linear regression. *J. Classification*, **5**, 249-282.

DeSarbo, W.S., Oliver, R.L. and Rangaswamy, A. (1989) A simulated annealing methodology for clusterwise linear regression. *Psychometrika,* **54**, 707–736.

Duda R.O., Hart, P.E. and Stork, D.G. (2000) *Pattern Classification.* Wiley, New York.

Fisher, D.H. (1987) Knowledge acquisition via incremental conceptual clustering. *Machine Learning,* **2**, 139–172.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA.

García-Escudero, L. A., Gordaliza, A. and Matrán, C. (2003) Trimming tools in exploratory data analysis. *J. Computat. Graphical Statist.,* **12**, 434-449.

García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008). A general trimming approach to robust clustering. To appear in *Ann. Statist.*. Technical report available at `http://www.eio.uva.es/inves/grupos/representaciones/trTCLUST.pdf`.

Gordaliza, A. (1991) Best approximations to random variables based on trimming procedures. *J. Approx. Theory,* **64**, 162-180.

Hartigan, J.A. (1975) *Clustering Algorithms.* Wiley, New York.

Hartigan, J.A. and Wong, M.A. (1979) A k-means clustering algorithm. *Applied Stat.*, **28**, 100–108.

Hastie, T. and Stueltze, W. (1989) Principal curves. *J. Amer. Statist. Assoc.*, **84**, 502-516.

Hennig, C. and Christlieb, N. (2002) Validating visual clusters in large datasets: fixed point clusters of spectral features. *Comput. Statit. Data Anal.,* **40**, 723-739.

Hennig, C. (2003) Clusters, outliers and regression: fixed point clusters. *J. Multiv. Anal.*, **83**, 183-212.

Hosmer, D.W. Jr. (1974) Maximun Likelihood estimates of the parameters of a mixture of two regression lines. *Comm. Statist.*, **3**, 995-1006.

Jolion, J.-M., Meer, P. and Bataouche, S. (1991) Robust clustering with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 791-802.

Kamgar-Parsi, B., Kamgar-Parsi, B., and Wechsler, H. (1990) Simultaneous fitting of several planes to point sets using neural networks. *Computer Vision, Graphics and Image Processing,* **52**, 341–359.

Kaufman L. and Rousseeuw P.J. (1990) *Finding Groups in Data.* Wiley, New York.

Lenstra, A.K., Lenstra, J.K., Rinnooy Kan, A.H.G., and Wansbeek, T.J. (1982) Two lines least squares. *Ann. Discrete Math.*, **16**, 201-211.

Maitra, R. (2001) Clustering massive data sets with applications in software metrics and tomography. *Technometrics*, **43**, 336-346.

McQueen, J. (1967) Some methods for classification and analysis of multivariate observations. 5^{th} *Berkeley Symposium on Mathematics, Statistics, and Probability.* Vol **1**, 281-298.

Meer, P., Mintz, D., Rosenfeld, A., and Kim, D. Y. (1991) Robust regression methods in computer vision: a review. *International Journal of Computer Vision*, **6**, 59-70.

Müller, C.H. and Garlipp, T. (2005) Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *J. Multiv. Anal.*, **92**, 359-385.

Murtagh, F. (2002) Clustering in massive data sets. In *Handbook of Massive Data Sets* (eds J. Abello, P.M. Pardalos, and M.G.C. Resende), Kluwer, 401–545.

Murtagh, F. and Raftery, A.E. (1984) Fitting straight lines to point patterns. *Pattern Recognition*, **17**, 479-483.

Ng, R.T. and Han, J. (1994) Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th Conference on Very Large Databases* (eds J.B. Bocca, M. Jarke, and C. Zaniolo), Morgan Kaufmann, 144–155.

Peña, D., Rodríguez, J. and Tiao, G.C. (2003) Identifying mixtures of regression equations by the SAR procedure. In *Bayesian Statistics 7* (eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West), Oxford University Press, 327–347.

Phillips, T.-Y. and Rosenfeld, A. (1988) An ISODATA algorithm for straight line fitting. *Pattern Recognition Letters,* **7**, 291–297.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Rousseeuw, P. J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.

Scott, D.W. (1992) *Multivariate Density Estimation.* Wiley, New York.

Silverman, B.W. (1986) *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Späth, H. (1982) A fast algorithm for clusterwise linear regression. *Computing,* **29**, 175–181.

Stanford, D.C. and Raftery, A.E. (2000). Finding curvilinear features in Spatial point patterns: Principal Curve Clustering with Noise. *IEEE Trans. Pattern Recognition*, **22**, 601-609.

Stewart, C.V. (1995) MINPRAN: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, 925-938.

Stewart, C.V. (1999) Robust Parameter Estimation in Computer Vision. *SIAM Review*, **41**, 513-537

Swayne, D.F., Cook, D. and Buja, A. (1998). XGobi: interactive dynamic data visualization in the X Window system. *J. Computat. Graphical Statist.,* **7**, 113-130.

Swayne, D.F., Temple-Lang, D., Buja, A. and Cook, D. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Comput. Statit. Data Anal.,* **43**, 423-444.

Tarpey, T. (1999) Self-consistency algorithms. *J. Computat. Graphical Statist.,*, **8**, 889-905.

Tarpey, T. and Flury, B. (1995) Self-consistency: a fundamental concept in Statistics. *Statist. Sci.*, **11**, 229-243.

Van Aelst, S., Wang, X., Zamar, R. H. and Zhu, R. (2006) Linear grouping using orthogonal regression. *Comput. Statit. Data Anal.*, **50**, 1287-1312.

Zhang, T., Ramakrishnan, R. and Livny, M. (1997) BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.,* **1**, 141–182.